A SNOMED Analysis of Three Years' Accessioned Cases (40,124) of a Surgical Pathology Department: Implications for Pathology-Based Demographic Studies

Jules J. Berman, Ph.D., M.D., G. William Moore, M.D., Ph.D., William H. Donnelly, M.D., James K. Massey, M.S.E.E. and Brian Craig
Veterans Administration Medical Center, University of Maryland School of Medicine, The Johns Hopkins Medical Institutions, Baltimore, MD; Shands Hospital and University of Florida, Gainesville, FL

ABSTRACT

Pathology departments devote considerable energy toward indexing diagnoses. To date, there have been no detailed tabulations of the results of these efforts. We have thoroughly analyzed three years' surgical pathology reports (40,124) generated for 29,127 different patients from the University of Florida at Gainesville between Jan 1, 1990, and December 31, 1992. 64,921 SNOMED code entries (averaging 1.6 codes per specimen and 1.4 specimens per patient) were accounted for by 1,998 distinct SNOMED morphologies. A mere 21 entities accounted for 50% of the morphology code occurrences. 265 entities accounted for 90% of the morphology code occurrences, indicating that the diagnostic efforts of pathology departments are contained within a small fraction of the many thousands of morphologic entities available in the SNOMED nomenclature. One of the key problems in using SNOMED data collected from surgical pathology reports is the redundancy of lesions reported for single patients (i.e., a patient's disease may be coded on more than one specimen from the patient, leading to false conclusions regarding the incidence of disease in the population). In this study, redundant SNOMED data was removed by eliminating repeat morphology/topography pairs whenever they occur for a single patient. SNOMED data can be stratified on the basis of age and sex (data fields included on every surgical pathology report). This analysis represents the first published analysis of SNOMED data from a large pathology service, and demonstrates how SNOMED data can be compiled in a form that preserves patient privacy.

INTRODUCTION

Before the advent of computerized laboratory information systems, pathologists were severely limited in the way they could obtain information related to the scope of their activities. Paper filing systems permitted pathologists to review the reports issued for a specific patient, but there was no practical way of summarizing data collected from many different patients. In the past, when pathologists were asked to comment on the incidence or age distribution of a lesion, at best they might quote a published statistic (from a report reflecting the experience of another hospital in another geographic and social environment) or offer a vague recollection from their own experience, such as, "I've seen half a dozen of these things, and they seem to occur in older people".

Despite the fact that modern pathology information systems all index reports under retrievable and universally recognized diagnostic codes (e.g., the International Classification of Disease (ICD)[1], or the Systematized Nomenclature of Medicine (SNOMED))[2], few services take the step of analyzing their own surgical pathology data. The reason for this is simple. Just like paper filing systems, modern laboratory information systems are only designed to answer queries related to a particular patient or diagnostic category. No laboratory information system supports unrestrained queries relating all report data fields and all diagnostic categories for all patients. Such an undertaking would consume considerable computational resources of the institution, would require additional programming effort and would provide a service of no direct clinical necessity to any specific patient.

Perhaps the most telling indicator of the difficulties associated with analyzing surgical pathology databases resides in the absence of published reports of organized global data summaries encompassing all the diagnostic entities encountered in the catchment population. The lack of such studies underscores the failure of pathology departments to satisfy the intended goals of indexed coding. According to Cote and Robboy, current systems of disease nomenclature and classification are directly descended from earlier classifications (beginning with the London Bills of Mortality in the early 1700's) created to determine the prevalence of diseases in a population [3]. Cote and Robboy, both principals in the development of SNOMED, suggest that a coding system should serve the needs of the entire health care system and provide data for epidemiologic studies and medical audit [3].

We have analyzed three years' SNOMED coded data obtained from a general hospital in Florida, eliminating diagnostic redundancies in the database and stratifying data based on age. This study addresses several important issues: 1) it demonstrates that the obstacles that must be overcome when preparing a database summary from raw data retrieved from the electronic files of a laboratory information system; 2) it offers a sample database to illustrate the values and limitations of SNOMED data and serves as a baseline for comparison with databases from other pathology services; and 3) it provides a way of preparing a complete demographic profile of the pathology received in a large hospital.

0195-4210/94/$5.00

MATERIALS AND METHODS

We examined data from 40,124 cases accessioned at the Shands Hospital, Gainesville, FL, between January 1, 1990, and December 31, 1992, inclusive. From these, there were 29,127 patients with complete demographics and 304 patients with incompletely coded reports, for a total of 28,823 patients with complete reports. Shands Hospital is a general teaching hospital for the University of Florida College of Medicine in Gainesville, Florida, which covers all major areas of medicine and surgery. Consultation cases were primarily referrals for oncology patients.

Approximately 90% of cases were coded by pathology residents, the remainder by faculty members. All coders participated in a two-hour tutorial course on SNOMED coding, taught by one of us (WHD). All coding was performed by referring to publications of the College of American Pathologists (CAP) that list the SNOMED codes [2], sometimes referred to as SNOMED-II, currently the most widely used edition of SNOMED. As a rule, each accession received one topography code and one morphology code per specimen. Redundant coding (assigning more than one morphology code to a specimen) was performed only for special cases, such as unusual tumors. On a daily basis, the pathologist enters terms into the various SNOMED fields. Although six SNOMED axes are accessible to the pathologist, the axes used at Shands Hospital are topography, morphology and procedure. From our own collected experience and from discussions with other pathologists, we feel that this is a very typical way of preparing SNOMED data. About 30 minutes was devoted each day to coding reports.

The computer used for the present study was an IBM PC/AT-compatible computer programmed with American National Standard M (ISO 11456 previously MUMPS), and the public-domain File Manager (FileMan) database management system of the United States Department of Veterans Affairs, used routinely in 169 VA medical centers [4]. Reports were obtained as a raw ASCII file of the M global variable that contained all the textual material and data fields for every surgical pathology report downloaded from the mainframe computer at the Shands Hospital, and containing the complete text of surgical pathology reports obtained between January 1, 1990, and December 31, 1992. The entire contents of each report, including patient demographics, date and time of accessioning and signout, specimen source, gross description, final microscopic diagnosis, pathologist's identification, and manually-entered SNOMED codes, were passed into the ASCII file, a total of 24 Megabytes. All routines were written with MGlobal (Houston, TX) M.

RESULTS

The distribution frequency of patients by age is shown in Table 1. The average age of patients who contributed tissue to surgical pathology was 35.8 years. The ability to stratify the ages of the population is extremely important, as it permits comparison of the data to other data sets for which the age distributions of the individuals are known (i.e. age adjustment).

TABLE 1. AGE DISTRIBUTION OF PATIENTS CONTRIBUTING SURGICAL PATHOLOGY MATERIAL

0-10 years old	3,096
10-20 years old	2,596
20-30 years old	5,038
30-40 years old	4,578
40-50 years old	3,301
50-60 years old	2,881
60-70 years old	3,958
70-80 years old	2,971
80-90 years old	665
>90 years old	43

One of the most difficult problems in extracting epidemiologically useful data from a SNOMED database is data redundancy. For instance, a single patient may have many basal cell carcinomas of the skin removed from various skin sites. A simple count of coded specimens may provide a false impression of the prevalence of basal cell carcinoma in the population. For epidemiologic purposes, the total number of people with basal cell carcinoma would, in general, be more useful than the total number of basal cell carcinoma specimens. The frequency distributions of the number of specimens submitted per patient is shown in Table 2.

Among the patients who had tissue submitted to pathology, there were, on average, 1.37 specimens per patient. The greatest number of specimens submitted for any patient in the 3-year study period was 21.

The total number of morphology codes in the database is 64,921. Redundant codes for patients were eliminated by preparing a list of all of the topography and morphology codes for each patient and eliminating topography-morphology pairs that shared the same first two digits of their morphology codes. The reason for matching only the first two digit-pairs was to allow for differences among pathologists in their choice of a morphology code (i.e., idiosyncratic differences in the last three digits). Considering the example of basal cell carcinomas in the patient population, the tumors may all have different topography codes (skin of face T02120, skin of neck T02300, skin of forearm T02630, etc.) and they may have different morphology codes (basal cell carcinoma M80903, morphea type basal cell carcinoma M80923, basosquamous carcinoma M80943) But for this example, any of the topography/morphology code-pair permutations deriving from the different topography and morphology listings will have the same pair of 2-digit leading strings (in this case T02/M80). Using matches in the first 2 digits of

topography/morphology code pairs effectively catches most redundancies due to coding idiosyncracy. Code idiosyncracy is a commonly occurring phenomenon [5]. It occurs when the same lesion is coded differently by different coders (e.g. one coder's basal cell carcinoma is another coder's basosquamous carcinoma). After elimination of redundancies (defined as two or more topography/morphology pairs identical to the first 2 digits of code) there were a total of 58,712 topography/morphology pairs. The ability to perform this elimination reliably is an essential step in SNOMED database interpretation.

TABLE 2. FREQUENCY DISTRIBUTION, NUMBER OF SPECIMENS SUBMITTED PER PATIENT

Specimens submitted	number of patients with the specified number of submitted specimens
1	22206
2	4378
3	1318
4	462
5	186
6	90
7	56
8	18
9	20
10	19
11	16
12	10
13	6
14	13
15	4
16	9
17	4
18	1
19	3
20	3
21	1
TOTAL	28,823

An interesting finding was that a very small number of morphologic entities account for the majority of morphology and topography codes. As shown in Tables 3 and 4, the 'median morphology code' (i.e. the 50-percentile morphology code representing the halfway point in the morphology code ranking) for manual coding occurs at rank 21. This means that at least 50% of all morphology codes are covered by the 21 most frequent (i.e., highest-ranking) diagnoses. 90% of all manual morphology codes are covered by the 265 most frequent diagnoses.

Table 4 shows a distribution of the 21 most common morphologies and their occurrences, ranked in descending frequency of occurrence, and accounting for 50% of all diagnoses made in the period of study. Non-diagnostic and non-specific morphologic codes account for the bulk of the high-frequency morphologies (e.g. normal tissue, no evidence of malignancy, inflammation).

TABLE 3. SUMMARY OF CODED MORPHOLOGIES FOR 40,124 SPECIMENS ACCESSIONED BETWEEN JAN 1, 1990 AND DEC 31, 1992

Total number of morphology codes	64,921
Number of disease entities accounting for 50% of the coded morphologies	21
Number of disease entities accounting for 90% of the coded morphologies	265
Number of disease entities accounting for 100% of the coded morphologies	1998
Average number of coded morphologies per accessioned specimen	1.6
Entities coded only once in the accession period	865

TABLE 4. LIST OF 21 ENTITIES ACCOUNTING FOR 50% OF ALL MORPHOLOGY CODES

	Number of cases
Normal tissue morphology	8712
Acute and chronic inflammation	2797
Chronic inflammation	2542
No evidence of malignancy	1774
Acute inflammation	1745
Adenocarcinoma	1441
Condyloma acuminatum	1315
Squamous cell carcinoma	1314
Protein Deposition	1193
Fibrosis	1063
Inflammation	968
Necrosis	882
Basal cell hyperplasia	871
Calcium deposition	864
Edema	716
Mild dysplasia	658
Products of conception	628
Proliferative Endometrium	588
Ulcer	587
Severe dysplasia	550

As shown in Table 5, the 'median topography code' (i.e. the 50-percentile morphology code representing the halfway point in the morphology code ranking) for manual coding occurs at rank 24. This means that at least 50% of all manual morphology codes are covered by the 24 most frequent (i.e., highest-ranking)

topographic locations. 90% of all topography codes are covered by the 213 most frequent sites.

TABLE 5. SUMMARY OF CODED TOPOGRAPHIES FOR 40,124 SPECIMENS ACCESSIONED BETWEEN JAN 1, 1990 AND DEC 31, 1992

Total number of topography codes	64,921
Number of anatomic sites accounting for 50% of the coded topographies	24
Number of anatomic sites accounting for 90% of the coded topographies	213
Number of anatomic sites occurring once only	933
Number of anatomic sites accounting for 100% of the coded topographies	1554
Average number of coded topographies per accessioned specimen	1.6
Number of uniquely coded entities (entities coded only once in the accession period)	621

The distribution frequencies for any topographic code or for any leading string of topographic code could be assessed by age or by sex or both. Table 6 is an example of the age distribution of all pancreatic neoplastic lesions encountered in the 3 year period of study. A pancreas topography code was considered to be any topographic code that began with the two-digit numeric string 59... This would capture T59000, Pancreas N.O.S. (not otherwise specified), as well as head of pancreas (T59100), pancreatic duct (T59010), etc. Just as Table 6 demonstrates the age distribution for all pancreatic lesions, a similar distribution could be achieved for lesions of any specified morphology code or leading numeric string of morphologic codes. All pancreatic neoplastic morphology codes were accounted for by 8 sets of 2-digit leading strings (M80, M81, M82, M83, M84, M88, M89 and M93). A table could be compiled that lists the age/sex distribution for all lesions of all topographic sites, but a single topographic site was selected due to limitations of space.

TABLE 6. DISTRIBUTION FREQUENCY OF ALL PANCREATIC NEOPLASTIC LESIONS BY AGE

<10	0
10 - 19	1
20 - 29	3
30 - 39	4
40 - 49	8
50 - 59	4
60 - 69	16
70 - 79	11
80 - 89	1
> 90	0

DISCUSSION

A pathologist's understanding of the incidence of diseases is determined by how often a lesion is encountered. This frame of reference is inherently biased and can lead to misleading impressions. For instance, in the 3-year database of the Shands Hospital, there were 415 hernia sacs and 26 cases of hemorrhoids. Hemorrhoids occur much more frequently than inguinal hernias, but a pathologist's experience would indicate otherwise. Actually, surgery is almost always performed for inguinal hernias, whereas patients with hemorrhoids seldom seek surgical relief. Thus, we should not use a surgical pathology database to determine the relative incidences of diagnosis or treatment may not involve surgery. Surgical pathology databases are good sources of data pertaining to lesions that must have biopsy confirmation or surgical treatment. We can probably get a reasonably good idea of the incidence of clinically-detected hernias in the patient population, because 1) a hernia repair is a general surgical procedure performed at virtually every medical center (i.e., patients do not cluster toward a few facilities that specialize in hernia repair); 2) a procedure is performed on the majority of patients with an inguinal hernia; and 3) tissue is received on almost every hernia repair.

Another error that results from estimating disease incidence by frequency of encounter relates to the multiplicity of biopsies associated with a disease process in a single patient. For instance, a single patient with chronic lymphocytic leukemia (CLL) may, over a period of several years, have the SNOMED morphology for CLL entered when a blood smear is assessed, when a lymph node is biopsied, when a skin infiltrate is sampled, when a spleen is removed, etc. For this reason, any analysis of disease frequencies must be able to represent data in a form where repeat morphologies for a patient are eliminated. In this study, redundant specimens for a patient were eliminated by searching for repeated topography/morphology code-pairs listed for a patient. However, this solution to the problem of specimen redundancy has its own drawbacks and may not be appropriate for all types of studies. For instance, patients may develop separate lesions of the same morphologic

type over a period of time (e.g., bilateral breast cancer), and an epidemiologist interested in this phenomenon may need to account for both tumors in a valid analysis of the incidence of cancers occurring in a population. Partly as a result of these difficulties, commercial laboratory information systems do not lend themselves to direct epidemiologic analysis, and database queries must be carefully designed to produce useful results.

In an effort to insure that diagnoses can be retrieved from databases, a variety of coding systems have been developed, all with the intention of categorizing disease entities as a unique number. Thus, a renal cell carcinoma, which may appear on a report as renal cell adenocarcinoma, hypernephroma, clear cell carcinoma, kidney carcinoma, kidney adenocarcinoma, adenocarcinoma of kidney or even as Grawitz tumor, can all be coded under the same, unique morphology and topography codes. Reports written in English, French, German, or any language, may all use the same code numbers to index their reports. Unfortunately, coding efforts may vary greatly in their accuracy. The reliability of indexed data related to diagnosis has received very little discussion in the medical literature. Hall and Lemoine, in one of the few available studies, found errors in more than 10% of indexed codes [5]. Currently, many pathology departments have employed automatic coding software and thus relieved themselves from the time-consuming burden of manual coding. In a recent study, we have shown that accurate automatic coding can only be achieved by monitoring the quality of the coded output and adding appropriate changes in the code look-up dictionary and in the manner that reports are written [6]. Furthermore, automatic coding can potentially produce databases with codes chosen in a uniform and predictable way optimized to support epidemiologic studies [6].

In the current study, the Shands Hospital laboratory information system was used only as the source of raw data files, not as a database engine supporting queries. Commercial laboratory information systems cannot budget their computational resources (the amount of computer time required to respond to a query) to perform in-depth database analyses. It is our observation that departments desiring full query access to their databases must acquire a devoted database application and then query their raw database file with their own programs written in a database specific language or a generalized database language (e.g., SQL, System Query Language).

Using routines written in the M programming environment, we have shown that it SNOMED databases can be fully analyzed, that the problem of code redundancy can be overcome, and that data relating the frequency of SNOMED morphology and topography entries according to patient demographics (age and sex) can be performed. SNOMED databases are one of the fastest growing and comprehensive medical databases, in that all U.S. hosptitals seeking accreditation by the College of American Pathologists or the Joint Commission for Accreditation of Healthcare Organizations must index all surgical pathology cases. In the last decade, most of the medical centers that had previously indexed their cases using card filing systems, have switched to electronic coding. SNOMED (specifically SNOMED version II) is, in our estimation, the most commonly used surgical pathology indexing system. A formidable amount of SNOMED data is accruing daily, and it would be a terrible waste if these data were not shared and analyzed. Unlike tumor registry data, which only provide cancer statistics, the SNOMED databases produced by surgical pathology departments cover every aspect of medicine. Prepared in the manner described in this study, SNOMED data can be tabulated as listings of topography and morphology codes, devoid of patient identifiers. Each record in a distributable database might consist of: 1) a unique patient identifier number that can be linked to a specific patient name by the contributing medical center only; 2) a list of topography and morphology code-pairs that describe all the different lesions biopsied for the patient exclusive of lesion redundancies; 3) the date of birth of the patient and 4) the sex of the patient.

REFERENCES

[1]. The International Classification of Diseases, 9th Revision: ICD-9CM, Second Edition, U.S. Department of Health and Human Services, Public Health Service, Health Care Financing Administration, U.S. Government Printing Office, 1980.

[2]. College of American Pathologists. Systematized nomenclature of medicine (SNOMED). College of American Pathologists, Skokie, 1976.

[3]. Cote RA, Robboy S: Progress in Medical Information Management: systematized nomenclature of medicine (SNOMED). JAMA 243:756, 1980

[4]. Davis R.G. FileMan: A User Manual. National Association of VA Physicians, Bethesda, 1987

[5]. Hall P.A., Lemoine N.R. Comparison of manual data coding errors in two hospitals. J Clin Pathol 39:622, 1986

[6]. Moore GW, Berman JJ: Performance analysis of manual and automated Systematized Nomenclature of Medicine (SNOMED) coding. Am J Clin Pathol 101:253, 1994